同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

交通数据分析

第六讲 空间数据：描述性分析

沈煜 博士 副教授
嘉定校区交通运输工程学院311室
yshen@tongji.edu.cn
2022年04月01日

计划进度

周	日期	主讲	内容	模块
1	2022.02.25	沈煜	概述	
2	2022.03.04	沈煜	在线数据采集方法	爬虫
3	2022.03.11	沈煜	线性回归模型	回归分析
5	2022.03.18	沈煜	广义线性回归	
4	2022.03.25	沈煜	广义线性回归（作业1）	
6	2022.04.01	沈煜	空间数据描述性分析	
7	2022.04.08	沈煜	空间自回归方法（作业2）	
8	2022.04.15	沈煜	关联：Apriori	机器学习
9	2022.04.22	沈煜	决策树、支持向量机（作业3）	
10	2022.04.29	沈煜	浅层神经网络	
11	2022.05.06	沈煜	卷积神经网络（期末大作业）	
12	2022.05.13	沈煜	经典网络结构	
13	2022.05.20	沈煜	聚类：K-Means、DBSCAN	
14	2022.05.27	沈煜	贝叶斯方法、卡尔曼滤波	
15	2022.06.03	-	端午节放假	
16	2022.06.10	沈煜	期末汇报（1）	
17	2022.06.17	沈煜	期末汇报（2）	

空间数据分析模块介绍

- 参考：MIT 11.S950
  - Applied Spatial Data Analysis with GeoDa/PySAL
- 基础：基本GIS操作与基本统计学知识
- 目标：空间分析的基本概念与空间回归基础
- 工具：
  - 基本工具：GeoDa与GeoDaSpace
  - 高级工具：基于Python的PySAL库

空间数据分析模块介绍

- **Luc Anselin**
- 空间计量经济学（Spatial Econometrics）理论奠基者之一
- 现芝加哥大学Stein-Freiler Distinguished Service Professor of Sociology
- 谷歌H指数93
- 总引用数8万+

主要内容

- **空间分析的基本概念**
- **空间自相关**
  - **空间自相关性统计**
  - **空间权重矩阵**
- 空间回归模型
  - 空间滞后模型
  - 空间误差模型
  - 模型选择
- 可视化

同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

基本概念

空间分析

- 空间分析的概念：
  - 对空间数据的属性和位置进行转换、操作与应用的一些列分析方法。（Goodchild et al）
- 空间分析研究的问题：
  - Where：在哪里发生？
  - Why：为什么在这些地方发生？
  - How：如何影响其他因素？如何被其他因素影响？
- 什么情况下需要进行空间分析？
  - 位置属性起到重要作用：当位置变化的时候，数据的信息内容随之变化。

空间数据分析的主要研究内容

- 地理信息映射与可视化
  - 展示
- 探索性空间数据分析
  - 探索
- 空间建模
  - 解释
- 空间数据科学的数据来源
  - 无所不在的采集与感知设备
  - 开放数据
  - 社交媒体数据（如地理标签）

空间数据科学

- 空间统计学
- 空间计量经济学
- 机器学习
- 计算机仿真

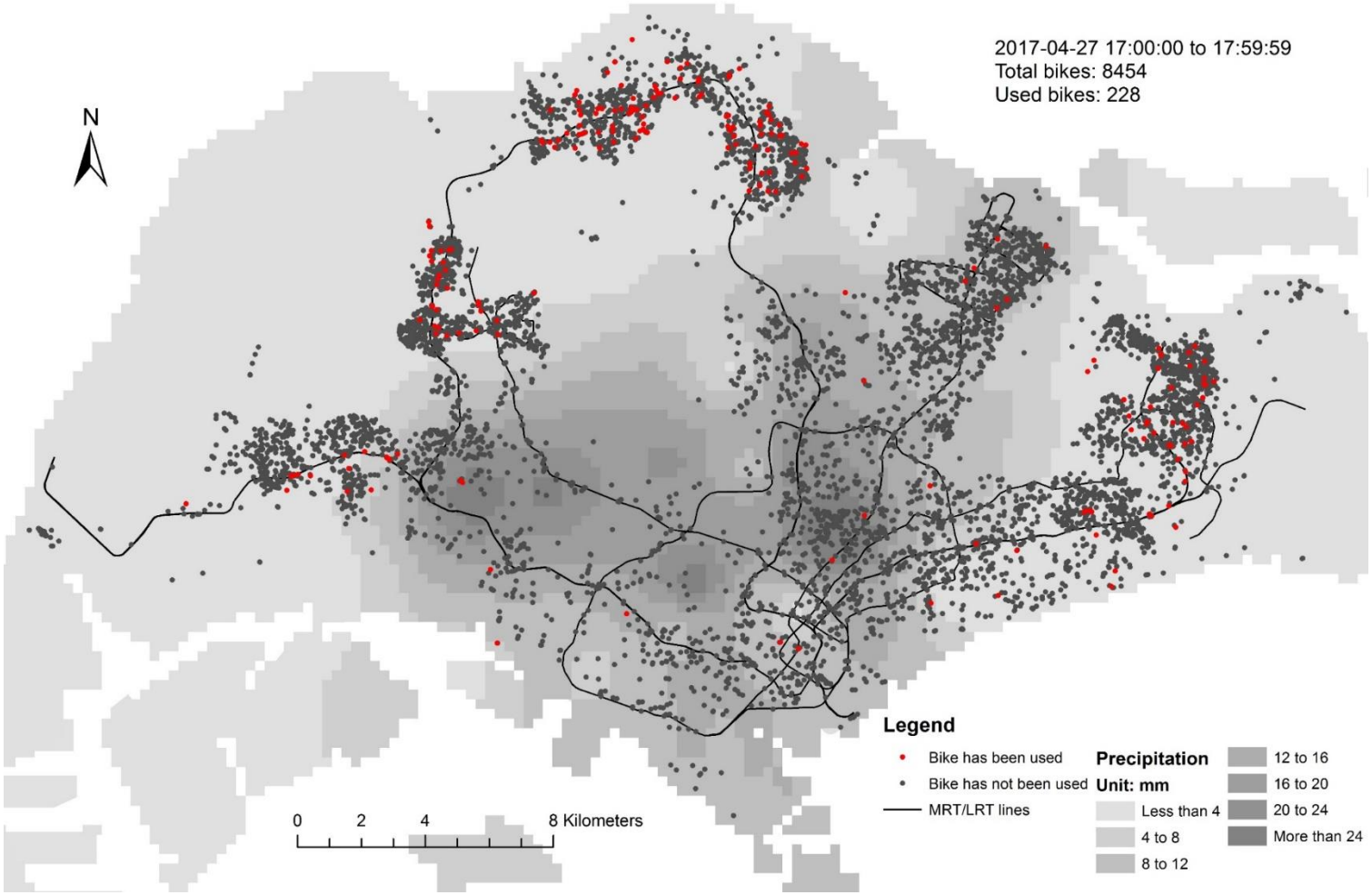
数据科学涉及到的内容

- 数据处理
- 数据融合
- 数据探索、模式识别、关联
- 可视化
- 建模（统计推断）、仿真、优化
- 借助软件工具
  - R, GeoDa等

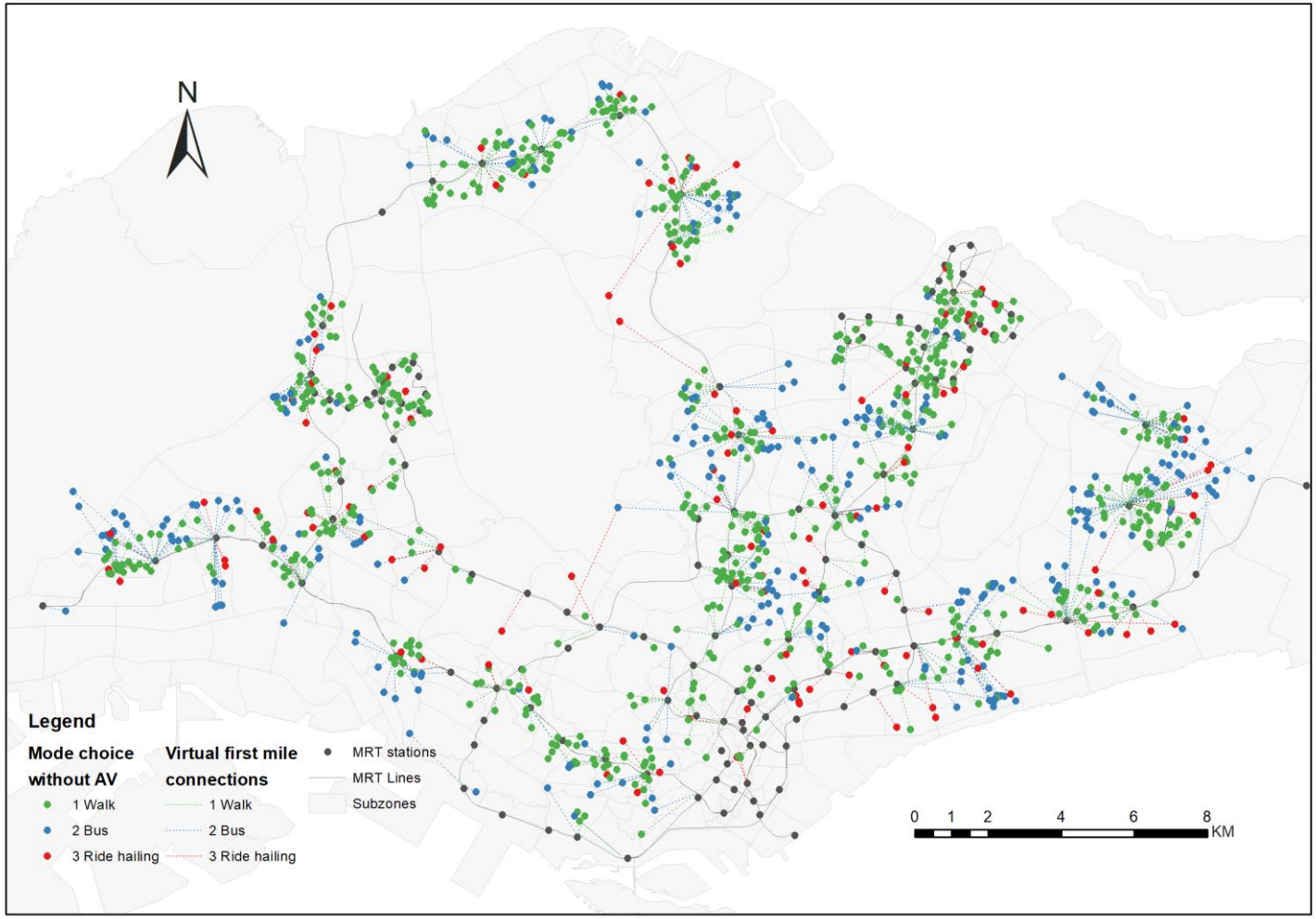
主要空间数据类型

- 事件（点）：如犯罪发生的地点
- 平面信息：如某地空气质量
- 地区数据：如行政区

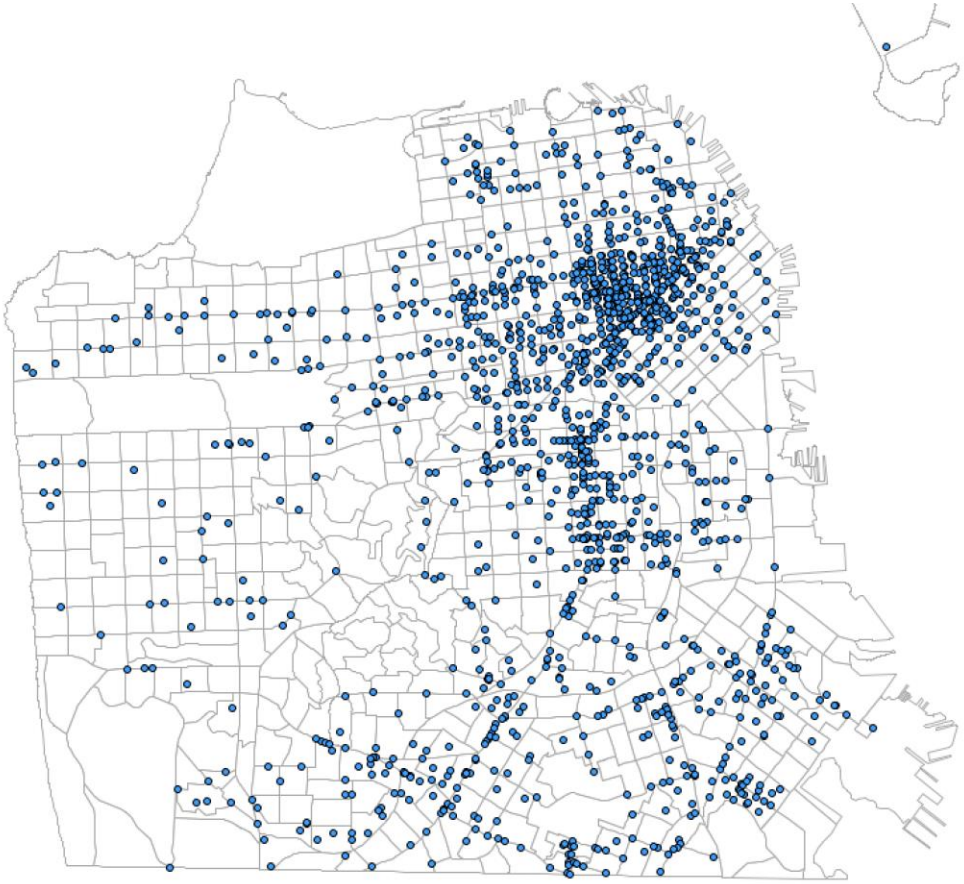
空间数据


2017-04-27 17:00:00 to 17:59:59
Total bikes: 8454
Used bikes: 228
N
0 2 4 8 Kilometers
Legend
Bike has been used
Bike has not been used
MRT/LRT lines
Precipitation
Unit: mm
Less than 4
4 to 8
8 to 12
12 to 16
16 to 20
20 to 24
More than 24

空间数据


N
Legend
Mode choice without AV
1 Walk
2 Bus
3 Ride hailing
Virtual first mile connections
1 Walk
2 Bus
3 Ride hailing
MRT stations
MRT Lines
Subzones
0 1 2 4 6 8
KM

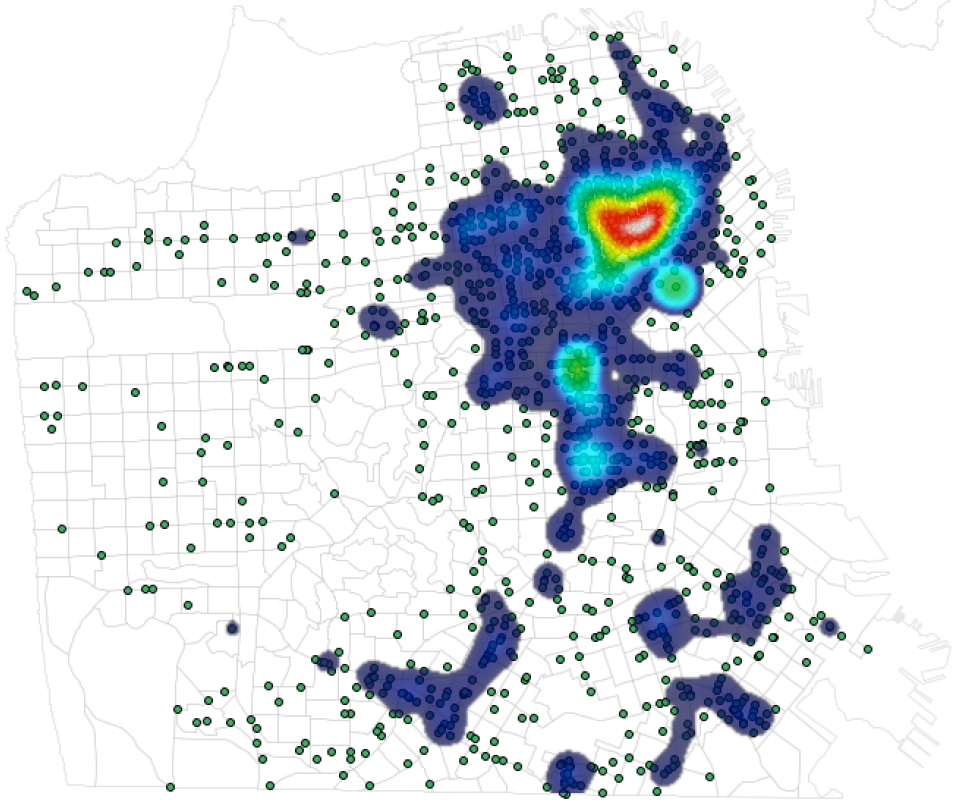
事件：旧金山车辆盗窃分布



对于事件（点）的分析

- 位置：所有事件，而非样本
- 研究问题：
  - 发生地点是否随机
  - 聚集的还是离散的
  - 聚集的位置

热力图



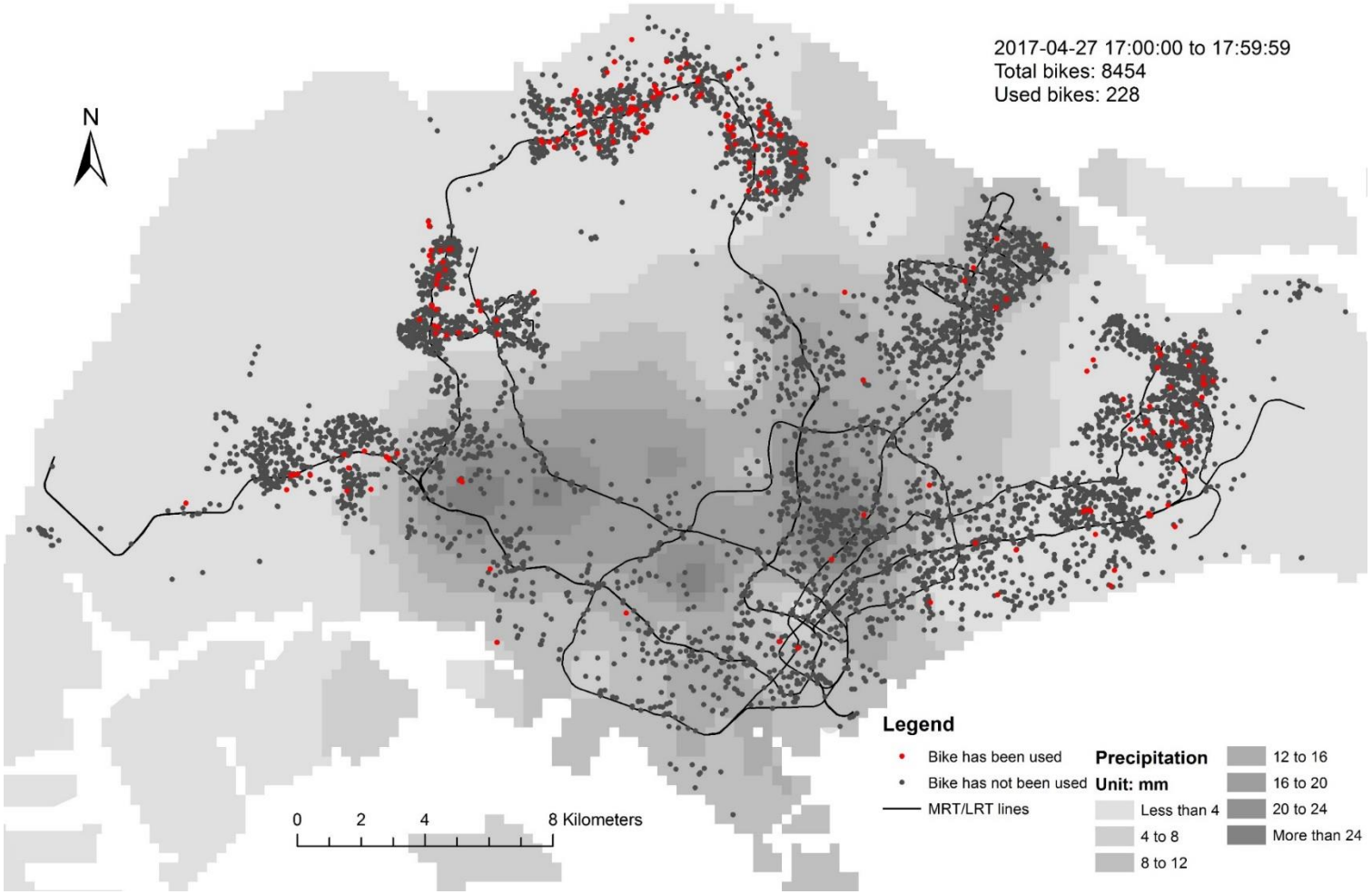
事件和点的关系

- 事件=点，但是并非所有的点都代表事件
- 有些点需要用来通过地理统计插值生成平面信息
  - 如降雨观测站
- 有些点需要用来代表固定地理区域
  - 如行政区的中心点

平面信息分析

- 以点信息作为采样点
- 将点上的信息填充（插值）至整个平面
  - 地理统计，克里金插值

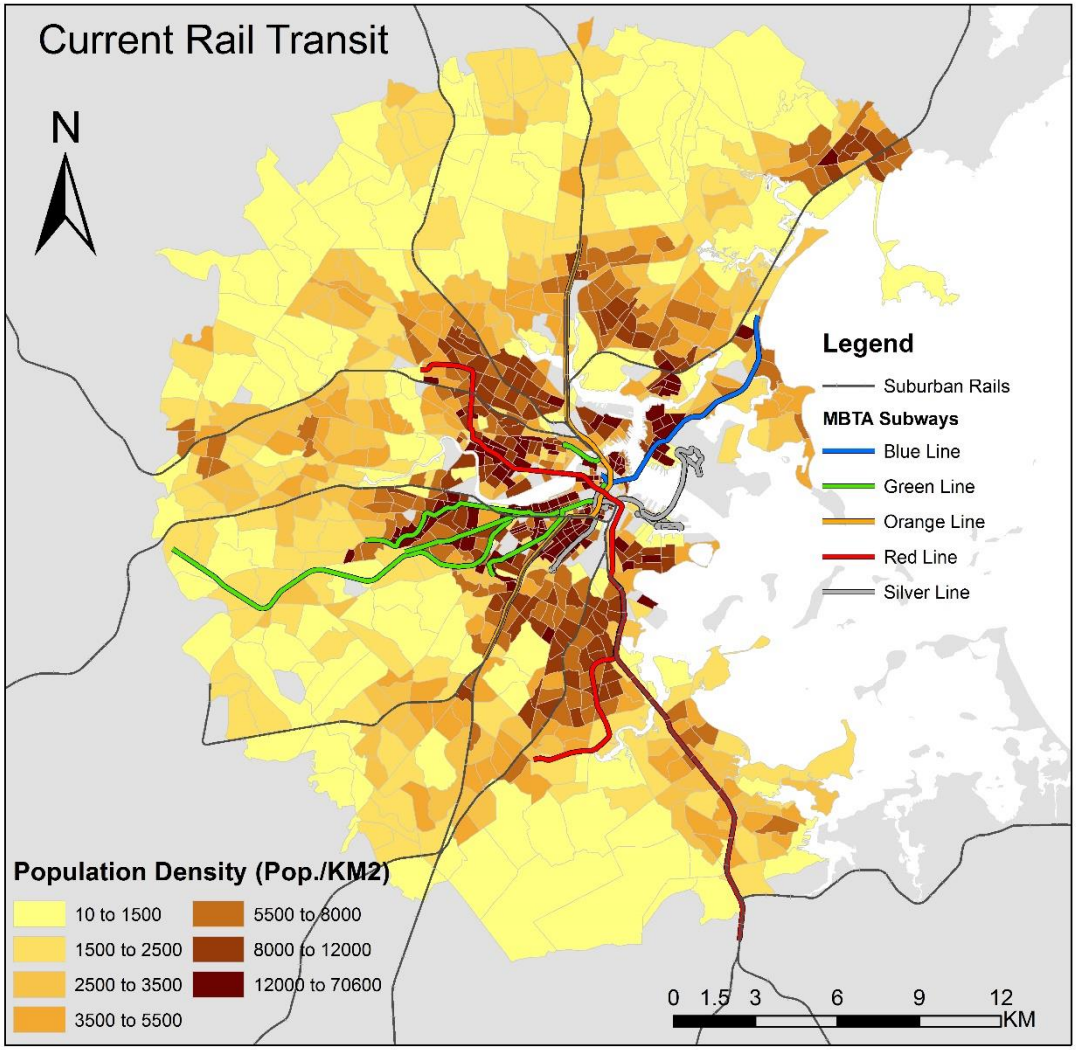
插值


2017-04-27 17:00:00 to 17:59:59
Total bikes: 8454
Used bikes: 228
N
0 2 4 8 Kilometers
Legend
Bike has been used
Bike has not been used
MRT/LRT lines
Precipitation
Unit: mm
Less than 4
4 to 8
8 to 12
12 to 16
16 to 20
20 to 24
More than 24

地区（或网格）数据

- 离散的地理信息
- 区域内所有的信息已知（不需要插值）
  - 如行政区，区域内的人口
- 研究不同区域数据的分布规律
- 相似的数据值是否在邻近的区域
  - 空间自相关
- 影响空间分布规律的因素有哪些
  - 空间回归分析

波士顿地区人口密度分布



MAUP问题

- Modifiable Areal Unit Problem
- 空间单位的大小影响空间分析结果
  - 空间异质性
  - 空间单位的面积和分布都会造成影响

同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

空间自相关

空间自相关

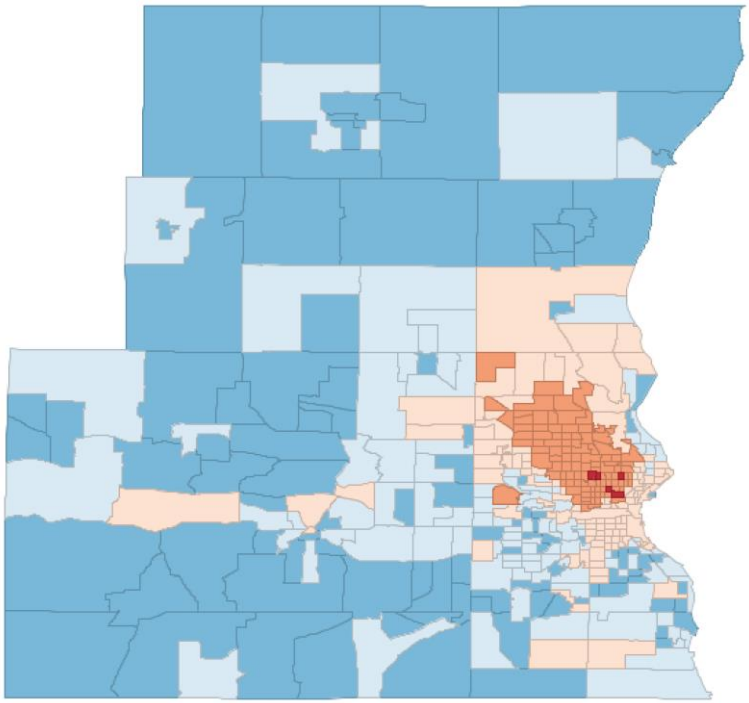
- 空间随机性
- 空间正相关、空间负相关
- 空间自相关统计

空间随机性

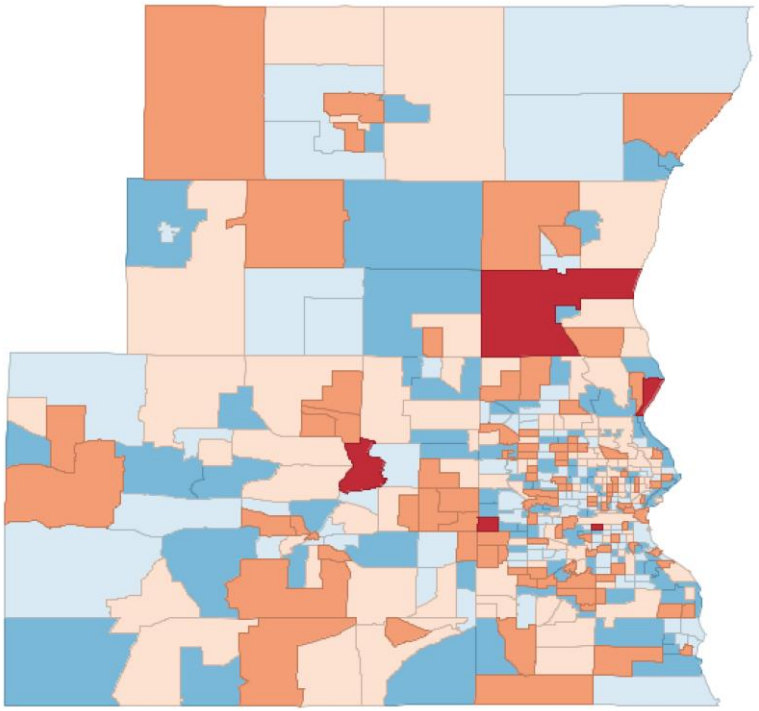
- 零假设
  - 空间分布没有任何规律
  - 如果拒绝零假设，则证明具有某种空间分布结构
- 解释
  - 观测到的分布规律和其他任何区域的分布都很类似
  - 在某地的数值并不依赖于其他（邻近）地方的数值
- 制造空间随机分布
  - 改变数据所在的位置
  - 不改变数据的信息

空间随机性

真实地图



生成随机分布

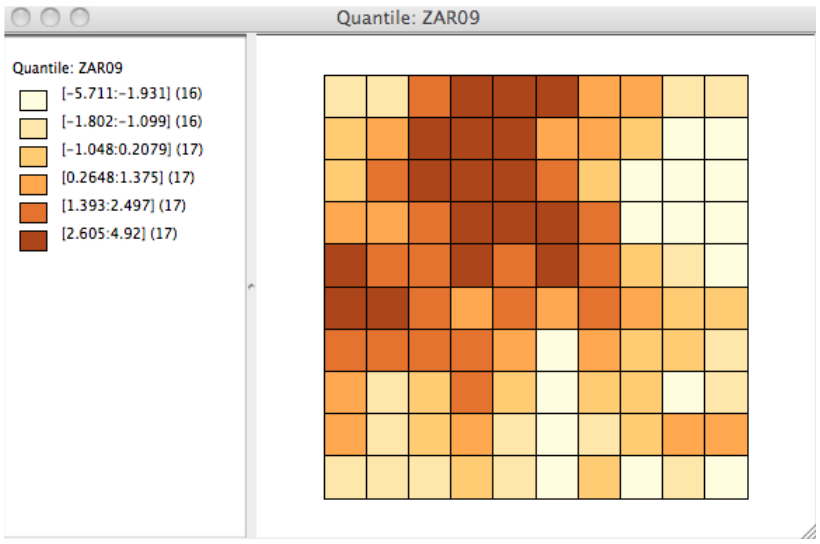


拒绝零假设

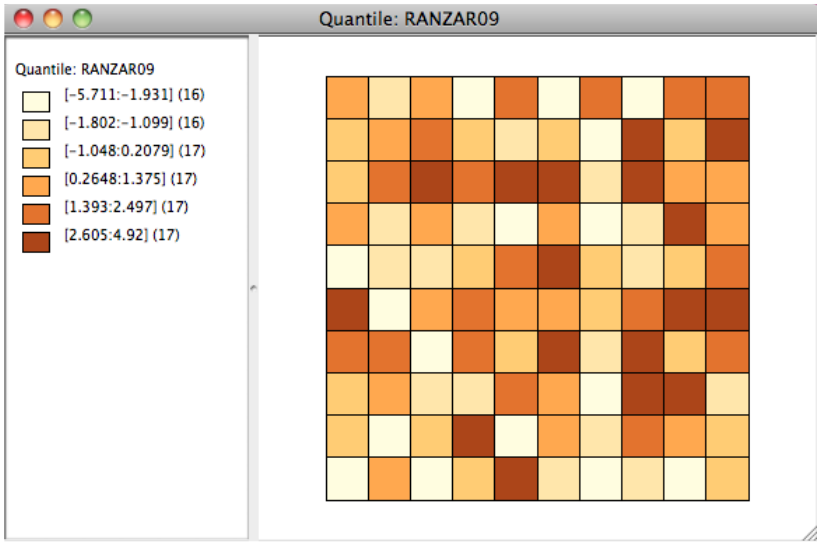
- 相似的数据在邻近地点间出现的频率高于空间随机分布
  - 空间正相关
- 相反的数据在邻近地点间出现的频率高于空间随机分布
  - 空间负相关

空间自相关：正相关

正相关

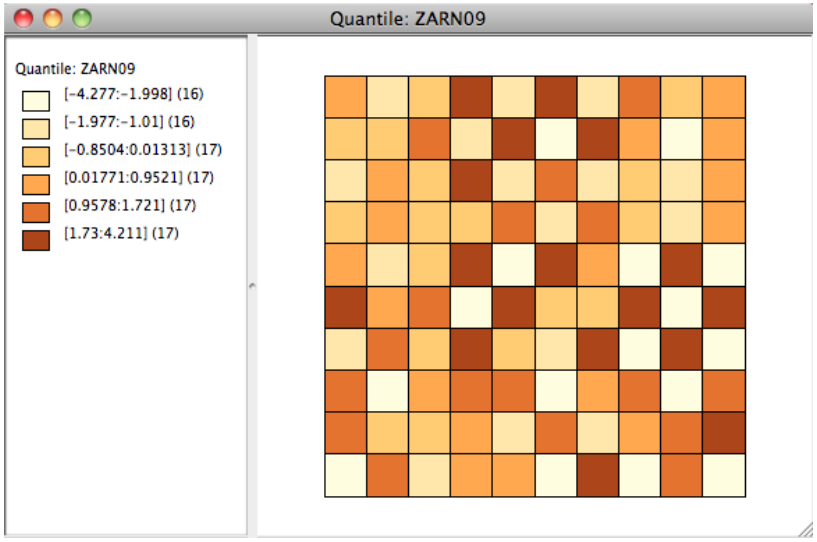


无关

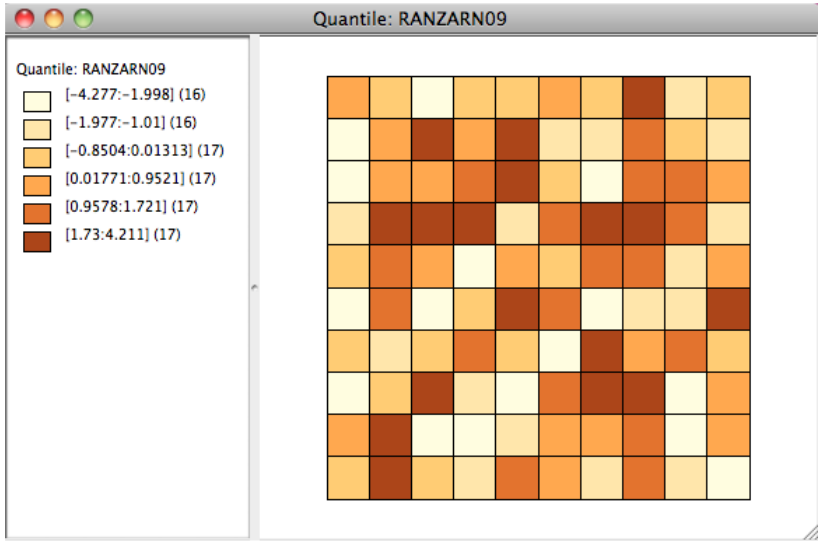


空间自相关：负相关

负相关



无关



空间自相关

- 正相关
  - 相似的数据聚集在一起
- 负相关
  - 分布规律类似国际象棋棋盘
  - 从视觉上很难与空间随机分布区分

同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

空间自相关性统计

统计检验

- 一个数值：对分布特征的总结
- 通过数据计算得出
- 统计检验：通过计算得出的结果与已知分布进行比较
- 有多大的可能性是零假设（空间随机）
- 零假设能否被拒绝

空间自相关统计参数

- 构造一个参数
  - 同时体现数值的相似程度与位置的邻近程度
- 体现不同位置上同一个观测变量的相似（或不同）的程度
  - 变量y
  - 不同位置i, j
  - 构造$f(i, j)$
- 相似：$y_i \times y_j$系统性增大或减小
- 不同：$|y_i - y_j|$或$(y_i - y_j)^2$系统性增大或减小

Moran's *I*

- Patrick Alfred Pierce Moran发明的空间自相关参数，是众多空间自相关参数中的一种
- $I = \frac{\sum_i \sum_j w_{ij} z_i z_j / S_0}{\sum_i z_i^2 / N}$
- 其中$z_i = y_i - m_x$（与均值的差）
- $S_0 = \sum_i \sum_j w_{ij}$
- 乘积$z_i z_j$与相关性参数的形式非常类似
- Moran's *I*的值随权重w_{ij}的变化而变化

Moran's *I*

- Moran's *I*的值通过S_0与N分别控制了分子与分母的尺度
  - S_0空间权重矩阵中的非零值或相邻对的数量
  - N观察的样本总量
- Moran's *I*的值在-1（负相关）到+1（正相关）之间
- 判断Moran's *I*是否显著与空间上的随机分布不同，需要同一系列随机生成的重排列数据进行比较

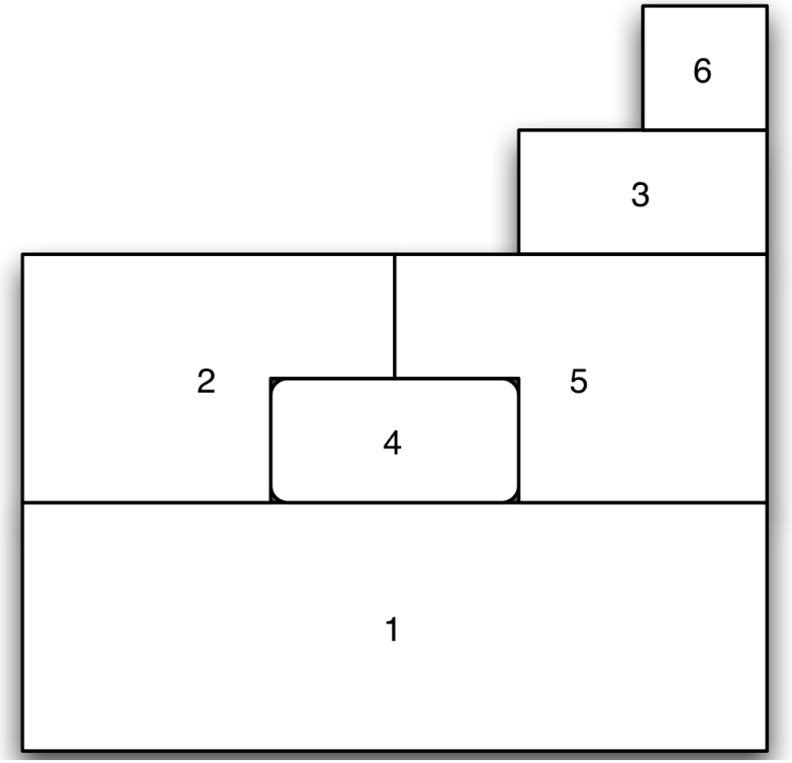
同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

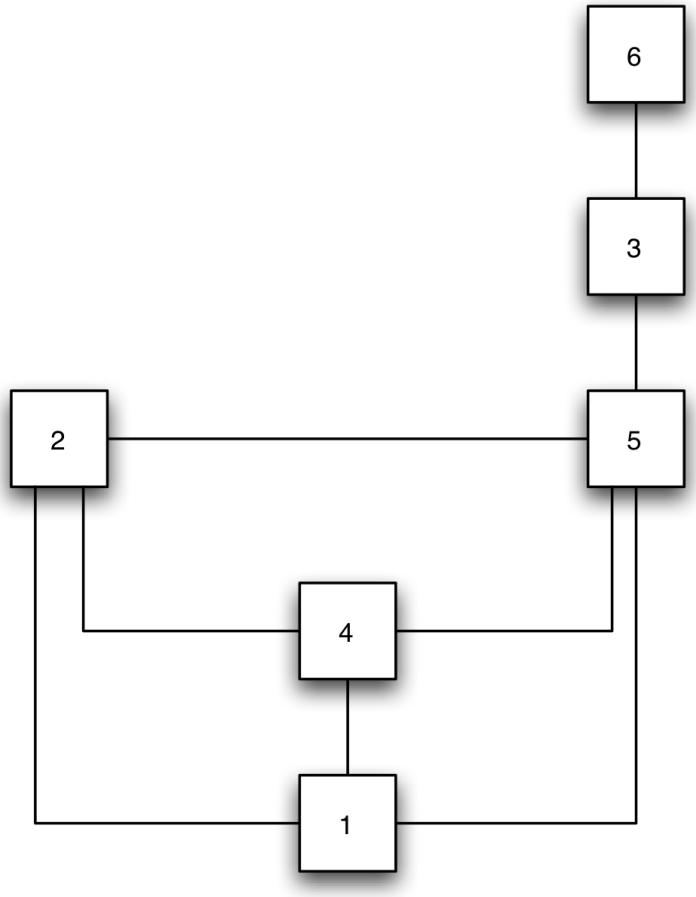
空间权重矩阵

空间权重矩阵

- 用来表达地区间的相邻
- 空间自相关研究空间上的交互
- 对所有n个地点，有$n \times (n-1)/2$对交互
  - 缺乏足够的信息来体现所有的交互情况
  - 每增加一个地点，增加的交互对呈平方式增长
- 解决办法
  - 排除一些交互，控制产生影响的地点的数量
    - 如只计算相邻
  - 用一个参数来体现，即空间自相关参数

定义共同边界才是相邻





构造空间权重矩阵

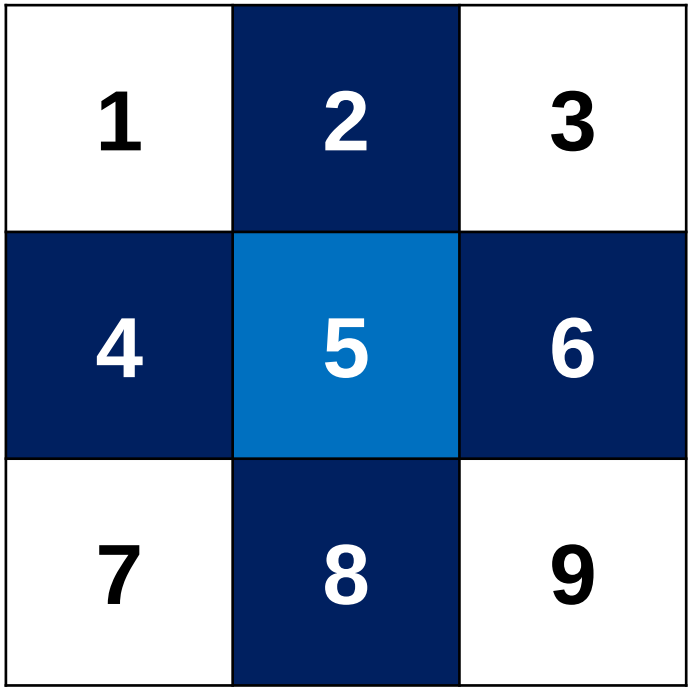
- 含元素w_{ij}的$N \times N$的正矩阵W
  - $W = \begin{bmatrix} w_{11} w_{12} & \cdots & w_{1n} \\ w_{21} w_{22} & & w_{2n} \\ \vdots & \ddots & \vdots \\ w_{n1} w_{n2} & \cdots & w_{nn} \end{bmatrix}$
  - $w_{ij} \neq 0$：相邻
  - $w_{ij} = 0$：不相邻
  - $w_{ii} = 0$：没有自相似性（self-similarity）

基于地理位置的空间权重矩阵

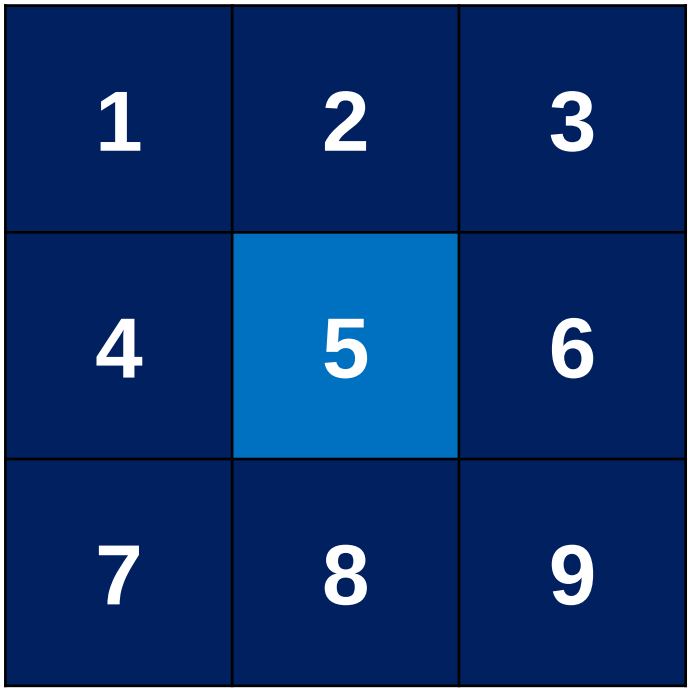
- 二元邻接（Contiguity）矩阵
  - 邻接（Contiguity）：共同边界
  - $w_{ij} = 1$：i和j有共同边界
  - $w_{ij} = 0$：i和j没有共同边界
  - $$W = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

几种主要相邻形式

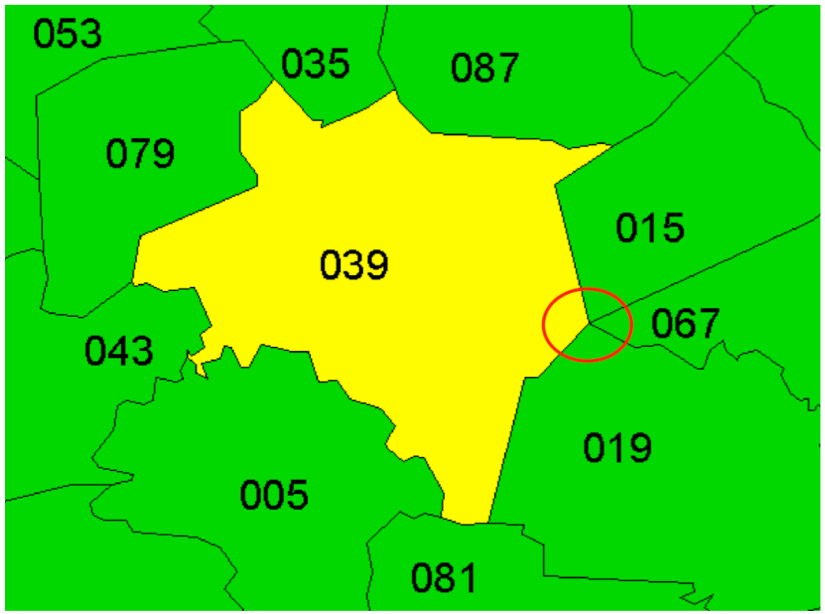
Rook Contiguity
共边



Queen Contiguity
共点或共边



几种主要相邻形式



- 039和067的相邻情况
- Rook contiguity
  - 不邻接
- Queen Contiguity
  - 邻接
  - 有共同顶点

基于距离的权重矩阵

- 距离
  - 点之间的距离
  - 多边形的中心点之间的距离
- 基于距离阈值d的权重
  - 对$d_{ij} < d$，w_{ij}为非零值
- k最近相邻矩阵
  - 每个观测数据具有相同数量的相邻数

对行进行标准化

- 变换使得$\sum_j w_{ij} = 1$:
- $w_{ij}^* = w_{ij} / \sum_j w_{ij}$
  - 控制参数变化的空间
  - 使得不同权重矩阵可比
  - 可以计算空间滞后性（spatial lag）
    - 邻居影响的平均

对行进行标准化后的权重矩阵

- $W^* = \begin{bmatrix} 0 & 1/3 & 0 & 1/3 & 1/3 & 0 \\ 1/3 & 0 & 0 & 1/3 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$

随机权重矩阵

- 行列双重标准化：$w_{ij}^{*} = w_{ij} / \sum_i \sum_j w_{ij}$
- 使得$\sum_i \sum_j w_{ij} = 1$
- 类似于概率
- $W^{*} = \begin{bmatrix} 0 & 1/16 & 0 & 1/16 & 1/16 & 0 \\ 1/16 & 0 & 0 & 1/16 & 1/16 & 0 \\ 0 & 0 & 0 & 0 & 1/16 & 1/16 \\ 1/16 & 1/16 & 0 & 0 & 1/16 & 0 \\ 1/16 & 1/16 & 1/16 & 1/16 & 0 & 0 \\ 0 & 0 & 1/16 & 0 & 0 & 0 \end{bmatrix}$

同济大学交通运输工程学院
COLLEGE OF TRANSPORTATION ENGINEERING
TONGJI UNIVERSITY

第六讲 结束